Multi Class Depression Detection Through Tweets using Artificial Intelligence

MUHAMMAD OSAMA NUSRAT*, National University of Computer and Emerging Science, Pakistan
WASEEM SHAHZAD, National University of Computer and Emerging Sciences, Pakistan
SAAD AHMED JAMAL, Université Bretagne Sud, France

Depression is a significant issue nowadays. As per the World Health Organization (WHO), in 2023, over 280 million individuals are grappling with depression. This is a huge number, and if this is; not taken seriously, these numbers will increase rapidly. Close to 4.89 billion individuals are social media users. People express their feelings and their emotions on platforms like Twitter, Facebook, Reddit, Instagram etc. These platforms contain valuable information which can be used to do research. Much work has been done on different social media platforms to predict depression. Still, there are some limitations in those works, such as binary classification, and incorrect labelling of the dataset. In this research work, we have used Twitter to predict five types of depression (Bipolar, major, psychotic, atypical, postpartum) in tweets based on lexicon labelling. We also used Explainable AI to give reasoning by highlighting the parts of tweets which represent that type of depression. Bidirectional Encoder Representations from Transformers (BERT) is used for feature extraction as well as for training. We used machine learning and deep learning methodologies to train the model. BERT delivered the most promising results, achieving an accuracy rate of 0.96.

CCS Concepts: • **Computing Methodologies** → **Artificial Intelligence**.

Additional Key Words and Phrases: Depression detection; Twitter Sentiment Analysis; Mental Health ; Text Classification; Machine Learning; Natural Language Processing; Social media and mental health ; Computational Linguistics

ACM Reference Format:
Muhammad Osama Nusrat, Waseem Shahzad, and Saad Ahmed Jamal. 2023. Multi Class Depression Detection Through Tweets using Artificial Intelligence. In . ACM, New York, NY, USA, 39 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Depression is a very serious mental disorder in which a person becomes hopeless and sad for a continuous period of time, depression is usually generated due to two different scenarios that occur with the subject, and both scenarios are quite uncommon; the first is the reason can be a situation from which the subject has extracted a traumatic experience and has retained unfondly memories from which the subject is unable to disconnect his or her thoughts and this situation, in turn, causes a long term retained depression. The second scenario is alienation caused by the subject's daily life events and experiences, which in turn causes the subject to lose motivation gradually, hence causing depressing thoughts over time, and those thoughts then transform into depression in the long run. The person experiencing depression tends to have no happiness; the things which a person used to enjoy at one time and were a source of happiness for him no

*Conceptualization, Muhammad Osama Nusrat, Waseem Shehzad; methodology, Muhammad Osama Nusrat, Waseem Shehzad; software, Muhammad Osama Nusrat, Saad Ahmed Jamal; validation, Waseem Shehzad; formal analysis, Muhammad Osama Nusrat, Waseem Shehzad; investigation, Muhammad Osama Nusrat; resources, Muhammad Osama Nusrat, Saad Ahmed Jamal; data curation, Muhammad Osama Nusrat, Saad Ahmed Jamal; writing—original draft preparation, Muhammad Osama Nusrat; writing—review and editing, Muhammad Osama Nusrat, Waseem Shehzad; visualization, Muhammad Osama Nusrat, Saad Ahmed Jamal; supervision, Waseem Shehzad; project administration, Waseem Shehzad. All authors have read and agreed to the published version of the manuscript.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Association for Computing Machinery.
Manuscript submitted to ACM

longer makes him happy. It not only rattles a person's mental health seriously, but it also affects the person physically, making him feel lazy; he may experience thoughts that he is useless,and he lives a purposeless life. Due to depression, a person's relationship is also damaged, and he lives in an abusive relationship. Depression could start or happen to a person if some significant incident happened, e.g. death of someone he was close to, bad marks in exams, a sudden injury, past regrets, etc. According to a report [35] in 2023, roughly 4.89 billion people will use social media on all platforms, like Twitter and Facebook. Instagram, Reddit, Pinterest, etc. This is an enormous number, and by this figure, we can imagine how important social media has become in our daily life. It is not wrong to say that it has become an integral part of our daily life.

The increase in the use of social media has also resulted in mental health issues [31]. Much research has been done in analyzing social media posts of users to check whether they are suffering from depression or not. In a research study, Fazida et al [20] proved that individuals who use social media have more chances of depression and anxiety. In the case of university students, when they see their assignments and homework and other activities that cause stress, they feel overwhelmed. To relax their minds, they scroll through social media, which helps them get instant gratification and a dopamine boost. But this is temporary because as soon as we think again of our assignment and the work we have to do, we again rush for a dopamine boost in the form of caffeine, binge-eating, watching movies, etc., and the cycle continues. This is why most people get depression, which is alarming.

According to the National Vital Statistics Report, Sally et al [12] did a study in which she aimed to compare the suicide rates of children and adults in America between the age of 10-24 years before Facebook launched from 2000 to 2017. According to the research, suicide rates were very low from 2000 to 2007, and they in and 57% from 2007 to 2017. According to Alexey Makarin, a professor at the Massachusetts Institute of Technology(MIT), Most students [37] reported being mentally sick after using social media. The main reason was peer comparison. Makarin and his team surveyed various universities to see Facebook's impact on student's mental health. The researchers concluded that the number of students feeling mentally distressed and ill increased compared to last year. Moreover, the number of students who were using Facebook in college for their anxiety disorders increased to 20%. Many students were reported to take antidepressants to cure their anxiety and mental illness.

There are several reasons and motivations for conducting research on sentiment analysis on social media platforms, i.e. Twitter. The first reason is that more than 4.89 billion people have active social media accounts, which is huge. The traditional methods like the Diagnostic and Statistical Manual of Mental Disorders (DSM), Patient Health Questionnaire (PHQ-9), Beck Depression Inventory (BDI), Hamilton Psychiatric Rating Scale (HRSD), and clinical interviews are costly and take time. Although they are the most effective, but social media sentiment analysis also has the capability to replace these effective methods in the near future. The reason why we chose Twitter as our focus platform for depression detection is that Twitter has a large amount of open-source data which is publicly available. By data, we mean tweets. Secondly, the data is in the form of text, which is easy to handle and maneuver around it. Moreover, all types of tweet data are available, which is fresh and old, which can help us to draw a comparison. Twitter alone has 396.5 million [40] active users, which is another reason why we chose it as a platform for performing sentiment analysis. Lastly, the biggest reason is privacy. Most of the users on other platforms like Instagram and Facebook post information that is not public on Twitter the information is public, which makes it a research-friendly platform. Facebook and Instagram mostly have image-based data. Although both of these platforms have a more significant number of users than Twitter, there is a drawback: it contains data in different languages, so it is a challenge to perform sentiment analysis in multiple languages. The research problem is to analyze tweets of users and predict whether a person using tweeter to tweet is depressed or not. Similarly, also predict the type of depression the user is facing. A considerable amount of work

has been done before to predict depression in tweets, but no work has been done to predict types of depression on Twitter. Jina et al. [22] worked on predicting the type of mental illness (anxiety. Bipolar disorder, borderline personality disorder, autism, and schizophrenia,) but they chose Reddit as their platform for research. My research problem is to predict the five most prevalent types of depression (Bipolar depression, Atypical depression, Psychotic depression, Major depressive disorder, and Postpartum depression). Moreover, we have used explainable AI to highlight parts of the tweet, due to which the model predicted the type of depression.

The first step is to scrap tweets with keywords or lexicons. We choose Apify, which is a scrapping platform. The lexicons which were used to scrap tweets were verified by the domain experts. After scrapping the tweets, preprocessing was done which is used to remove hashtags, URLs, @, punctuation removal, links, usernames, and stop words and converts the tweets to lowercase. The tweets which were not in English were excluded. Similarly, spammy tweets were also excluded. After that, tweets were labelled according to the lexicons. The tweets in which the person himself doing the tweet is not depressed is not marked as depressed. After labeling tweets, we did tokenization, splitting the data into train and test, and then using BERT for feature extraction and training. Finally, explainable AI is used to give reasoning if a certain tweet is marked as depressed, then what is the reason behind it.

Following is the significant contribution to our research.

- Dataset scrapping and Labeling
- Predicting types of depression from the tweets
- Multi-class classification of tweets
- Explainability

We have divided our work in five sections. In The first section we have discussed the introduction and background of the problem what is the motivation behind conducting this study and discussed our contribution. In the second section, we have discussed the literature review. We have also made a comparison table in which we have discussed each study its strength and weakness. We have also discussed the problem statement, research questions and inclusion and exclusion criteria. The third section discuses we have discussed research methodology, which includes pipeline discussion about the dataset, the language which we used to execute our problem. In the fourth section, we have explained our implementation and results. In the concluding section, we deliberated upon the final takeaways and potential future directions of this research.

2 RELATED WORK

Rafal et al. [29] participated in a competition organized by Codalab. The task was to predict intensity of depression in the tweets whether the person doing posting the tweet is severely depressed, moderately depressed or not depressed. The dataset was provided by the event organization. The researchers proposed their own solution in which they first fine-tuned BERT, RoBERTa and XLNet on the dataset which was provided to them. Among all the models, the RoBERTa-large model delivered the superior results. The researchers then took one step further and then again fine-tuned RoBERTa large on the dataset provided. The Reddit mental health dataset was also used alongside the dataset provided by Codalab, the competition's organizers. The team of authors and his fellow researchers won first prize in the competition.

Safa et al. [33] proposed a new method that can automatically collect a large number of tweets from users and then analyze them to see whether they contain depressive features or not. In the tweets which were collected, the people self-reported to be depressed. A multi-modal framework was also proposed, in which they took the help of the n-gram language model and Linguistic Inquiry and Word Count (LIWC) to predict depression. For, e.g. if a user tweet contains

a certain word such as feeling lonely, feeling sad etc., we can analyze it using the n-gram language model, similarly, by counting the frequency of words which came in a sentence using the LIWC dictionary, we can also check whether a post shows depressive content or not. For example, if a tweet frequently contains the word depressed, sad, or alone, there is a high probability it is depressed.

Zhao et al. [38] conducted research in new south wales to analyze people's tweets to check whether people feel more depressed after COVID-19 or not. He analyzed the tweets between Jan 2020 and May 2022. The researchers suggested a novel classification model centered around multimodal characteristics. The findings indicated an increase in depression among individuals following the COVID-19 pandemic. In this research, there was a strange trend, which was that people still remained worried even when the restrictions imposed by the government due to COVID-19 were removed. The main reason was that they were worried that due to relaxation, the spread of covid would increase further. These results will help the government to know that people need mental health help and assistance.

Rissola et al. [30] proposed a new method that can help collect a dataset of social media posts that contain depression or not. The author emphasized that due to the lack of dataset, there is much difficulty in building a model which can detect depression with good accuracy, so the dataset shared by the author has the capability to predict depression accurately. The researchers trained their dataset using BERT model and the results were very good with good accuracy, precision, recall and F1 scores. Further research can be done using this dataset, and it will also be useful for the mental health doctors. This new method of automatically collecting huge dataset will assist the current as well as upcoming researchers to build tools and apps which can predict depression accurately.

Jina Kim et al. [22] introduced a deep learning model that has the potential to determine mental health illnesses such as bipolar disorder, schizophrenia, autism, bipolar personality disorder, and autism. The author used Reddit to collect data for their research work. There are 6 classes in the dataset, the researchers faced a class imbalanced issue and to cope with that issue they used SMOTE algorithm. They used XG Boost classifier and CNN to classify Reddit posts.A separate binary classification model was made to classify each mental illness to improve accuracy as in some cases there are users who mention they are suffering from multiple type of mental illness which will make the model confuse, and it will not be able to efficiently classify the data. The evaluation metrics used were accuracy, f1 score, precision recall, etc.

Guntuku et al. [16] presented a study in which they discussed that we could also analyze whether a person is depressed by looking at Facebook or Twitter profiles images. The researchers chose 28749 users of Facebook who were suffering from depression and anxiety and used this dataset for training. Then the model was validated at 887 posts of users on Twitter. In the final step, the model was put to the test with 4132 distinct Twitter users. The people who posted depression-related content had their profile images with only faces of themselves they were not seen in any group photos, and profile images were grayscale, and they had low visual harmony.

Tadesse et al. [36] used NLP techniques to detect depression in Reddit posts. The researchers found a common term, also called a lexicon of terms, standard in Reddit posts of depressed people. The lexicon of terms was most used by depressed users in the posts. These lexicons, also called features, were found by applying machine learning and natural language techniques.The Bigram combined with SVM yielded an accuracy rate of 80% and an F1 score of 0.8 but the ensemble model gave the highest accuracy (i.e., LIWC +LDA+ bigram) of 91% and an f1 score of 0.93.

Islam et al. [19] used machine learning techniques to detect depression on Facebook. The people posting on Facebook sometimes express their feelings in emoji's, sometimes in the form of comments, so the first step was to extract features. For that purpose, authors used LIWC to extract the features from Facebook, whether a Facebook post or people's comments. The author used machine learning techniques to classify features extracted by law. These algorithms included

decision trees, SVM, KNN, and ensemble classifiers. The decision tree outperformed all other classifiers, standing as the best performer. Most of the depressive user's comments were posted from midnight to early morning (AM), while very few depressed comments were posted during daytime.

Ferwerda et al. [14], in their research, discussed that there is a relationship between the pictures and the personality traits of the person posting the picture. For that purpose, they focused on the Instagram platform and researched a sample of 193 Instagram profiles of users with their consent. They used Google vision API to gather the dataset of users' images on Instagram. There were a total of 54,962 images. k mean clustering machine learning algorithm was used to do clustering means the users who posted most photos with musical instruments love adventures and new experiences. They like exploring new things. These people are willing to explore new cultures. Similarly, people who post pictures related to clothing, sports, and fitness are very self-disciplined. These people have good work ethics. People who post pictures with some electronic instrument, e.g., a picture of themselves at a concert, have a personality trait of extraversion. Similarly, people who post pictures of themselves in fashionable clothing and who participate in extracurricular activities have a personality trait of agreeableness. This means these people are cooperative and supportive. Lastly, people who have fewer pictures with clothing and more pictures with jewelry or other materialistic thing have the personality trait of neuroticism means they are insecure low self-esteem, etc., so, in this study, the author discussed five personality traits of people by analyzing their pictures on Instagram using machine learning.

Chen et al.[7] discussed that NLP techniques had been used in detecting a specific type of depression.However, only a handful of studies have employed detailed sentiment analysis techniques to identify a person's mental health through their social media posts. So in this research, Chen and his fellow researchers used an emotive sentiment analysis algorithm that can extract fine-grained emotions from the tweets of persons. Emotive has nine features of emotions: sad, happy, disgust, shame, surprise, fear, confusion, anger, and an overall score. So this emotive algorithm gives fine-grained emotion scores to a tweet, indicating the tweet's dominant emotions. By fine-grained emotions, we mean the emotions people use while writing or in their speech. Using machine learning algorithms, these emotions were used as features to diagnose people with self-reported mental health conditions. SVM and random forest classifier showed the best results.

Manoj et al. [23, 27] discussed basic machine learning techniques for data along with their location information. They focused on detecting depression from tweets using NLP. First, tweets were scrapped, and the necessary preprocessing methods were applied. Next, a hybrid text embedding technique was used to convert text data to numbers, including fast text + TF-IDF (Term frequency Inverse Document Frequency). After that, classifiers based on machine learning were applied to the dataset to determine whether they contained depression. SVM and Random Forest classifiers were utilized, with the Random Forest classifier achieving the highest accuracy, at 75%.

Kumar et al [24] focused on the Twitter platform to gather tweets and analyze whether they contain depression. The tweets were collected using Twitter API. These were the raw tweets; after that, we applied preprocessing techniques such as hashtag removal, mentions, and url removal. NLTK library was used to tokenize each tweet. Frequent words like the, is, am, and are, which have no significance to determine depression, are also removed; then, the next task was to assign a lexicon score to each tweet. The tweets with the more depressive keywords will have a high lexicon score. The dataset was trained using SVM and a Naive Bayes classifier, with SVM achieving an accuracy rate of 93%.

Singh et al. [34] used ensemble learning to classify the tweets as moderately depressed, severely depressed, or not depressed. Codalab provided the dataset used for the competition. After preprocessing, the author fine-tuned BERT RoBERTa and xlnet for predicting the labels. Subsequently, the author implemented an ensemble voting classifier. Every

model will predict a classification for the tweets. The label receiving the most votes will be selected in the end, and the tweet will be labelled according to the highest vote. The accuracy was 0.6253, and the team of authors won 3rd prize.

Junyeop et al.[6] suggested a lexicon-oriented strategy for depression detection in Korean, English, and Japanese tweets. A lexicon means keywords or phrases people use when discussing depression in social media posts or tweets. The lexicon for depression was made in all three languages, and the psychiatrists later verified it. The data of users was collected using the Twitter API. Following that, a lexicon was employed to allocate a score to each tweet. Subsequently, the labeled dataset of tweets was used to train the machine learning model, this model is used to classify new tweets as depressed or not depressed.

Priyanka et al. [1] analyzed tweets of people suffering from depression and then applied machine learning algorithms to classify the tweets as depressed. The tweets were scrapped from Twitter, and then they were labelled. The criteria for labeling involved categorizing a tweet as "depressed" if it contained words such as "depression," "anxiety," or "mental illness," and as "not depressed" otherwise. There were a total of 3754 tweets. After doing preprocessing and feature extraction, the researchers used SVR(Support vector Regression) and multinomial naive Bayes algorithms for training and classification. Support vector regression gave the highest accuracy of 79.7%.

Aswathy et al.[2] proposed an app which can help in detection of depression. The main idea is that the user will enter several inputs to the system such as I am feeling depressed etc. If the sentence has depression features the app will tell that you have depression, else it will negate that you don't have depression. The main magic is inside the system and how it is working and how it is made. A dataset of tweets was used by the author, which was imbalanced 11911 tweets were normal, and 2308 tweets were depressed. An ensemble model of CNN & LSTM was used for training the dataset of tweets posted by the user on Twitter. The ensemble model results were far better than those using a simple SVM model. The ensemble model accuracy was 0.97, which was far better than the SVM accuracy of 0.83.

Bata et al. [4] in this study presented a new tool called Ardep, an Arabic lexicon that can identify lexicon in Arabic that people use when they are depressed. To make Ardep, a massive dataset of tweets in the Arabic language was gathered, and five psychiatrists verified the lexicon. Ardep has 5922 lexicons that indicate depression. Ardep is very valuable and can be used to find signs of depression present in the content shared on various social media platforms. of people in Arabic. It is an asset for mental health institutions in Arabic countries.

Glen et al. [11] participated in a hackathon called CLPsych(Computational Linguistics and Clinical Psychology). The hackathon held at John Hopkins University had the objective of identifying depression and post-traumatic stress disorder (PTSD) from tweets. In easy words, the participants have to analyze the tweets of the user to see whether they are suffering from depression or not and propose a model which can give high accuracy.

Paula et al.[8] discussed an app called Psychologist in a Pocket in this paper. The primary motivation behind making this app was to detect depression early. People use text messages, tweets, Facebook posts to express their feelings. The author and his fellows gather the words which indicate depression on social media. For that purpose, they took the help of mental health physicians and students in college. The reason for this was that most depression symptoms are found in teenagers in their high schools and universities because of studies pressure, insecurities, self-esteem issues, anxiety about paying loans and debts, etc., so after gathering information from students and health physicians, they were able to find a lexicon used by social media people experiencing depression and trauma. Anxiety etc., so this application can check whether a person doing text messages or posts has symptoms of depression.

Ginetta et al.[9] discussed how threatening depression can be for humans if left untreated. It affects all parts of the human body, including a person's mood and health. Doctors prescribe medications for depression-suffering individuals, such as serotonin, but in some cases, individuals have treatment-resistant depression, and these medicines have no

effect on their bodies. The percentage of people with this type of depression is 12-28%. These patients require high-dose medication to cure illnesses, like antipsychotics and electroconvulsive therapy.

Choudhary et al. [13] discussed in this research about Twitter and how it can be used to predict Major depressive depression (MDD) in individuals. In this research, crowdsourcing was used to find people suffering from depression. Crowdsourcing means, in this context, gathering data on individuals who are depressed. The data was gathered via surveys. The researchers concluded that people experiencing depression had specific traits, which were low interaction with people and highly sensitive. These features were gathered, and then a machine learning model was trained on these attributes to predict depression if a person has these features. SVM was used, and accuracy came to be 70%.

Lin et al. [26] proposed a depression detection app called Sense Mood, which can detect depression from textual and visual information on Twitter. Firstly, it finds and gathers textual and visual features which exhibit depression. After that, these features are combined to classify whether the individual posting the tweet is in a depressive state or not. Firstly, a dataset of tweets was gathered, i.e., the tweets which contain depression or no depression. Then, the textual and visual features were extracted using BERT and CNN. After combining the features, researchers classified the tweets using machine learning models.

Coppersmith et al. [10] first showed the importance of social media and how it can be used to predict whether a person is suffering from depression. The author of this paper introduces a novel method for collecting the dataset of different types of disorders like post-traumatic stress disorder (PTSD) and seasonal affective disorder (SAD). Then, the author proposed classifiers that can be used to detect each type of depression with reasonable accuracy.

Victor et al. [25] firstly emphasized that depression is more in rich countries, i.e., 90% of the well-off counties have high suicide rates in them. As previous studies have mentioned, social media is an available source of information and can be utilized to predict an individual mental state. The primary goal of the author in this paper was to diagnose depression at an early stage if an individual posts something online. If there is a sequence of depressive posts at a particular time, this will indicate that the person has a high chance of having depression. The contribution by the authors in this research was to use better feature extraction techniques which can extract textual features of the tweets better than previous approaches, improve the accuracy of the models and also apply a genetic algorithm and check whether it increases the results of the classification or not.

Akhtar et al. [18] in this research did a survey or a questionnaire among university students in Bangladesh to see if they were suffering from depression or not. Four hundred seventy-six students participated in the questionnaire, and the results were shocking, as 15 percent of students suffered from moderate to high depression. The reason for the depression was the tuition fee the students had to pay during the pandemic when they suffered through the crisis.

Kecojevic et al.[21] researched university students who were in their undergraduate degree to see the effects of COVID-19 on them. The research was conducted on students from a college in New Jersey, as New Jersey was the most affected area of COVID-19. The survey was done among 162 students in New Jersey. In the survey, some questions were asked to the students, including what difficulties they face in their life and in their studies. Most of the students were female in the questionnaire. Students also discussed that due to online classes, their studies were affected, and they could not concentrate. Some other factors were low wages. The author emphasized that the college authority should pay attention to the students and try to resolve their mental health problems.

Burnap et al. [5] developed a classifier that can help classify texts related to suicide and no suicide on Twitter. First, lexical features, which showed depression/suicide in the tweets, were extracted. Then the classifier was trained based on those lexical features. The classifier was able to detect suicide ideation in the tweets. The motivation of this research

was to build a classifier that can be used to identify posts of individuals who have suicidal ideation to help individuals who have suicidal thoughts. This will lower the risk of suicide.

Renata et al.[32] proposed a solution that can help psychiatrists and mental health physicians to detect if a person or patient has mental health issues. The person's mood can be checked by his social media posts, as most people depict and express their thoughts and feelings in their social media posts. So if we devise a model which can first depict depression and then alert the doctors or family members of the patient about the patient's critical situation of the patient, it will help save a person's life.

Ghosh et al. [15] presented a bidirectional LSTM CNN model with attention mechanisms to detect depression in the social media platform in the Bangla language. In previous studies, lexicon-based labelling was used for feature extraction. The attention mechanism is used in this research for feature extraction, as it focuses on relevant and essential parts of the text. The incorporation of attention mechanisms resulted in a notable improvement in the performance of the model. The accuracy came to 96 percent.

Sooji et al. [17] presented a hierarchical attention mechanism for depression detection on Twitter. Previously the main focus was to increase the classification accuracy of the machine learning classifier, but more work needs to be done on explainability. If a model is classifying a tweet to be depressed, then why is it doing so? So in this research, the model aims to highlight parts of the tweet, looking at which it is marked as depressed or not depressed.

Zogan et al. [39] proposed an explainable approach, called Multi-Aspect Depression Detection with Hierarchical Attention Network (MDHAN). Its purpose is to determine depression in the users' social media posts. It uses a hierarchical attention mechanism that can help find or extract essential features in the text data indicating depression. Hierarchical attention mechanisms can highlight significant words in a sentence and introductory sentences in a document. MDHAN combines demographic data, clinical data, and social media posts of the patient. After extracting features from all types of data and combining them, it gives attention scores to all the relevant parts. It highlights the features when predicting whether the input text reflects depression. In this way, mental health clinicians will be able to verify it, and they will be able to know if the model is making any mistakes.

Zucco et al. [40] told the need for sentiment analysis and how it can be helpful for depression detection in people. The researchers in this paper explained the applications of sentiment analysis in depression prediction. The researchers aimed to present a model to clinicians and mental health doctors which can tell the progress of a depression patient whether he is recovering with the passage of time or not. Furthermore, the authors also proposed a prototype for depression detection that uses multimodal features such as facial expressions, speech, language, etc.

Liang et al. [41] said we could use sentiment analysis to extract information about a user's opinion and thoughts. So we can also take the benefit of sentiment analysis in the field of medical science. Several deep learning algorithms have been used for sentiment analysis, but there needs to be more work done on explainability. It means that if a model predicts an output, what is its reason? Cutting-edge deep learning models are challenging to understand, so there is a need for explainable AI models.

Bacco et al. [3] highlight the importance of explainable AI models and why they must be used. These models also tell us that if they made a particular decision, why did they make that decision? This will develop more trust in humans and practitioners in AI and help them figure out biases, especially in healthcare and medicine. If we are making an app that can help diagnose depression in humans, there must be reasoning behind why it tells that a particular person is depressed. This will help psychiatrists to develop more trust in these systems, and in case of any mistake, they will correct it. The researchers also proposed an attention-based document classification and document summary system, which uses an attention mechanism to generate document summaries.

Table 1. Summary of Literature Review

Ref No,Year	**Summary**	**Strengths**	**Weakness**
[29], 2022	The task was to make a model which can classify tweets as severely depressed, moderately depressed, and not depressed. The researchers fine-tuned BERT, RoBERTa, and XLnet on the dataset and then used ensemble learning to predict depression in the tweet.	The paper achieved the highest accuracy among all other papers.	Accuracy can further be improved.
[33], 2022	The objective of the research was to identify depression in individuals based on their self-reported social media posts.	A multimodal framework extracted features from the textual and image data. It will capture information about features more comprehensively.	Self report depression diagnosis
[38], 2021	The study aimed to investigate the dynamics of community depression in New South Wales, Australia, as a consequence of the COVID-19 pandemic.	This study will help the government to know how much individuals are suffering from depression, and it will help the government to take steps to	The study was limited to only one city in Australia, which was new south wales. This study can be extended to multiple cities.
[30], 2020	The author put forward a technique that enables the automatic collection of tweets using the Twitter API.	The dataset is huge, which can be used to build a robust model.	Twitter API gives limited number of tweets.
[22], 2020	The authors introduced a deep-learning model capable of classifying different types of mental disorders (depression, anxiety, autism, mental disorders)	The deep learning model has the potential for utilization by mental health clinicians to detect depression in their patients.	In this research, the authors only considered Reddit as a social media platform; moreover, the researchers did not consider the sociodemographic and regional differences.

[16], 2019	The author presented a model which can predict depression from the user profile on Twitter.	This is one of the few models that can predict depression using user profiles.	The study was conducted on a low sample size of people.
[36], 2019	The researchers used Reddit as a platform to find a lexicon that depressed people use to express their emotions and then used an ensemble model of LDA, LIWC, and bigram with multilayer perceptron as a classifier.	The ensemble model achieved higher accuracy than the individual models.	Challenging problem not easy to implement
[19], 2018	The researchers used Facebook to predict depression from the Facebook posts and comments of the users. LIWC was used for feature extraction from Facebook comments and posts. Several machine learning classifiers were used, which included SVM and Decision Tree. The decision tree classifier outperformed other classifiers	Using multiple machine learning techniques helped the researchers that which technique is giving them good results.	The study was conducted on a single platform. Moreover, facebook has privacy concerns. Most of the users keep there data private.
[14], 2018	The researchers proved that there is a correlation between users' pictures on Instagram and their personality traits. With the consent of 193 users their Instagram profiles were analyzed using google vision API. There were a total of 54,962 images which were gathered.	The personality trait of a user can be predicted by looking at the images which they post on Instagram.	Accuracy is low.
[7], 2018	This research aimed to identify self-reported depression and then used five machine learning classifiers to predict depression.	EMOTIVE was used as a sentiment analysis algorithm in this research, extracting fine-grained emotional features from a person's tweets.	The data is based on self reported statements.
[23], 2022	In this paper, the researcher's main aim was to predict depression from the tweets posted by people on social media.	The model was able to predict depression from the input tweets.	Accuracy was low, i.e, 75%.

[24], 2021	The authors collected Twitter data through the Twitter API, performed data preprocessing, extracted linguistic features, and utilized machine learning algorithms for training and evaluating prediction models.	Lexicons are also called features. The tweets were labeled as depressed based on the lexicon verified by domain experts.	The study is limited to Twitter only.
[34], 2022	Ensemble learning was used to classify the tweets as moderately depressed, severely depressed, or not depressed. Codalab provided the dataset used for the competition. After preprocessing, the author fine-tuned BERT RoBERTa and XLNET for predicting the labels. After that, the author applied an ensemble voting classifier.	Ensemble voting gave high accuracy rather than using individual classifiers	The dataset is small accuracy is low.
[6], 2022	Proposed a lexicon-based approach to detect depression in tweets in multiple languages, which were Korean, English, and Japanese.	BERT gave the highest F1 score	It is only applicable to text data.
[1], 2019	Analyzed tweets of people suffering from depression and then applied machine learning algorithms to classify the tweets as depressed or not. The authors employed SVM (Support Vector Machine), Multinomial Naive Bayes, and Support Vector Regression algorithms in their study. Support vector regression gave the highest accuracy of 79.7%.	A novel method is discussed which is used to classify depressed tweets from non-depressed tweets.	The study is limited to binary classification. It can be extended to multiclass classification.
[2], 2019	used an ensemble model of CNN & LSTM for identifying depression in tweets posted by the user on Twitter. The ensemble model results were far better than those using a simple SVM model.	LSTM +SVM gave an accuracy of 85%.	The quality of the dataset significantly influences the outcomes. If the dataset is not of good quality, even ensemble models may not produce satisfactory results.

[4], 2019	A lexicon related to depression in Arabic was created so that it can be used to evaluate depression	Mental health physicians in Arab countries can benefit from this app.	This app can only benefit Arabic people.
[11], 2015	The author participated in a hackathon called CLPsych. The objective of the hackathon was to identify depression and post-traumatic stress disorder (PTSD) by analyzing tweets. In easy words, the participants have to analyze the tweets of the user to find whether they are suffering from depression or not and propose a model which can give high accuracy.	The researchers collaborated with each other, which was a positive sign.	It was only focused on predicting PTSD and depression.
[8], 2016	The researchers proposed an app that can predict depression based on the lexicon of words it has been trained.	This app can be used to prevent suicide by monitoring people's social media activities.	In the future, mEEG can also be added to the app, which will aid in detecting depression at a fast pace.
[9], 2018	The study investigates the effects of ketamine on structural plasticity in human dopaminergic neurons.	An important area of research has been explored which is structural plasticity.	The results of research cannot be generalized on all everyone.
[13], 2013	The author discussed that Twitter could be used for predicting major depression. Crowdsourcing was used to collect information about Twitter users who have been suffering from depression using the CESD square test.	A statistical classifier was made which can predict that a person has depression in its initial phase and he is likely to get depression if it is not controlled.	sample size was small.
[26], 2020	The author proposed a depression detection app called Sense Mood which has the capability to detect depression from textual and visual information on Twitter.	The accuracy of predicting depression is high because it extracted both textual and visual features.	Sometimes this app will fail to distinguish between true emotions and sarcasm.

[10], 2014	The author proposed a new method to collect datasets of various types of mental disorders and then classify them utilizing machine learning techniques. The data was collected from Twitter spanning the years 2008 to 2013.	Authors took the help of LIWC software and statistical software, which can help them identify the common language used by individuals who are suffering from depression. It helped them find patterns that are common in the language used by people with mental health disorder people.	Most other mental disorders, like binge eating and Alzheimer's, are rarely discussed on Twitter. Most people who have depression do not even reveal information on social media, which is also a problem.
[25], 2017	The primary goal of the author was to diagnose depression at an early stage of an individual posting something online. If there is a sequence of depressive posts at a particular time, this will indicate that the person has a high chance of having depression.	Better feature extraction techniques that can extract textual features of tweets.	There is still room for improving accuracy, we can develop a better classification system for our model.
[18], 2020	A survey was carried out among university students in Bangladesh to check whether they were suffering from anxiety and depression. If yes, why are they, and what is the reason and factors behind it.	Students who didn't exercise and do their homework on time suffered from depression more than the ones who did contrary.	The snowball strategy was used due to less time and limited resources, rather than choosing the random sample's strategy.
[21], 2020	A survey designed to assess the mental health of undergraduate students at New Jersey University, post covid.	This survey helped the university authority to take steps to improve the mental health of students at New Jersey university.	It was a self-reported survey, not verified by mental health physicians whether a student is really mentally upset or not.

[5], 2017	Classify suicide-related communication on Twitter using a multi-class machine classification approach.	The study offers a machine learning approach to identify suicide-related communication on Twitter, which can help with the early detection and prevention of suicide.	The study focuses only on Twitter data, which may not be representative of the general population.
[32], 2016	To develop a model which can predict whether a given text has suicidal ideations or not.	The authors were successful in making a classifier which can accurately distinguish whether a given text has suicide symptoms or not.	The research was focused on one platform only.
[15], 2023	The article presents a proposed model that utilizes attention mechanisms, bidirectional Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) to detect depressive texts in Bangla language on social media platforms.	Proposed bidirectional LSTM and CNN model gave high accuracy on dataset.	The researchers need to have more annotated Bangla text to increase accuracy. Moreover, a more diverse dataset can be made from different platforms like TikTok, Instagram, LinkedIn, Facebook, Pinterest, etc.
[17], 2022	A novel language analysis method involving metaphor concept mapping to identify and analyze how individuals with depression express their emotions and experiences.	Enhances understanding of emotional expression in depression through metaphor analysis.	The researchers plan to conduct a large-scale study that will focus on categorizing different characteristics of depression. To do this, they will analyze the metaphorical and cognitive expressions used by users on social media to describe their experiences with depression.

[39], 2022	The researchers put forth a deep learning model that combines multiple aspects and features to enable the explainable detection of depression on social media platforms.	The model proposed in the study achieved higher accuracy compared to existing state-of-the-art approaches.	The evaluation of the model was conducted on a limited dataset and focused solely on detecting depression in English-language social media. As a result, its generalizability to other languages and contexts may be constrained.
[41], 2018	The hybrid approach suggested in the research combines rule-based techniques and machine learning methods to enhance both the interpretability and precision of sentiment analysis specifically in the medical field.	Explainable AI has been used where the model is also explaining about the decision it makes.	The study only focuses on sentiment analysis in the medical domain and may not be applicable to other domains or contexts.
[3], 2021	The researchers proposed two transformer-based architectures for sentiment analysis classification. Additionally, they incorporated an extractive summary to provide an explanation for the model's decision-making process.	The model was able to achieve state-of-the-art results.	Computational cost is high.

2.1 Research Gap

From above table, we have found some limitations, which are that most of the work which has been done so far in the field of depression detection is based on binary classification, i.e. A social media post has depression or not. Very few studies [22] have executed multi-class classification, and on Twitter, multi-class classification is not yet explored. We have addressed this gap in my research by predicting the five most prevalent types of depression, which are bipolar depression, atypical depression, psychotic depression, major depressive disorder, and postpartum depression. Similarly, very few studies have used explainable AI in their models for reasoning. This gap is also addressed in my research. We have used explainability that if the model has predicted that a certain tweet has bipolar depression or some other type of depression, then it will highlight the words in that tweet (i.e. these are the words due to which the model predicted this type of depression) The dataset labeling was also a crucial part as in most of the above studies if a social media post has the word depression, then that post is labeled as depressed; similarly, if a tweet has the word depressed, it is marked as depressed, which is not true in most cases. The dataset was labelled by considering the context of the whole sentence, which is also a research gap.

2.2 Problem Statement

Social media platforms contain tons of valuable information which can be used for sentiment analysis. We have used Twitter to predict types of depression (Bipolar, Psychotic, Atypical, Postpartum, and Major depressive disorder) in the tweets. Moreover, we have also used explainable AI to give the reasoning for each prediction of the model.

2.3 Research Questions

Below are the research questions for my thesis.

- Which type of keyword or phrases are used by people suffering from a particular type of depression?
- Can we use machine learning and deep learning techniques to measure types of depression accurately?
- Can NLP accurately predict types of depression in tweets?

2.4 Inclusion and Exclusion Criteria

Below are the inclusion and exclusion criteria for our research.

2.4.1 Inclusion Criteria.

- The people who have self-reported being depressed currently and in the past are included in this study
- The tweets which contain the lexicons "I have bipolar depression," "I am suffering from atypical depression," and "I have a major depressive disorder" and vice versa are also included.

2.4.2 Exclusion Criteria.

- The tweets which are not in the English language are excluded.
- Spammy (which contains only hashtags) and repetitive tweets are also excluded.
- Retweets are also excluded.
- The incomplete tweets or tweets whose sentences are not complete are also excluded.

3 RESEARCH METHODOLOGY

According to the American psychiatric association, the term "depression" refers to a "syndrome that is characterized by a clinically significant disturbance in an individual's cognitive abilities, emotional regulation or behavioral patterns. The different surveys reported that 20% of people of all ages face some form of mental illness at some point, with approximately eight percent of adults having had severe depression. Aside from the severity of mental disorders and their impact on an individual's psychological and physical health, social stigma or discrimination has caused individuals to be neglected by the community and to avoid taking the necessary treatments.

This is in addition to the fact that mental disorders impact an individual's psychological and physical health, as literature has demonstrated the inherent challenges of diagnosing mental disease through social media platforms. Numerous researchers have attempted to discover crucial findings through various natural language processing methods. This presented the intrinsic difficulties of diagnosing mental problems such as depression. Acquiring adequate knowledge about the specific field of research is necessary to successfully create an accurate predictive model and extract the most prominent features in the data [28]. Even if these features were removed, this does not guarantee that those characteristics are the primary contributors to achieving improved accuracies can be obtained. Because of these factors, we are investigating the prospect of utilizing deep neural networks because the features are learned within the design. The explanation of each step of the methodology is explained in subsequent sections.

3.1 Proposed Pipeline

Millions of people all around the world are coping with the incapacitating effect of depression, which is a widespread mental health issue. Although various treatments are available for depression, many people who suffer from the condition do not get the care they require. This may be because they cannot recognize the signs of their sadness or do not have access to mental health services. In recent years, researchers have turned to social media platforms like Twitter as a potential data source for detecting individuals experiencing depression. Specifically, they have been looking for user tweet patterns that may indicate they are depressed. For this, research endeavors proposed different techniques using Natural Language Processing and Machine Learning. Among them, the BERT-based techniques provide promising results. These machine learning and Natural Language processing techniques have significant repercussions for the general health of the population, as they can facilitate the identification of individuals at risk of developing depression and subsequent linking of those persons with appropriate mental health resources.

These NLP and machine learning-based techniques are multistep processes that must be completed to detect the accurate depression from the tweets. The first thing that needs to be done is to compile a dataset of tweets and classify them according to whether or not they show signs of depression. This can be a strenuous effort to do because depression is a complicated and varied disorder that can present itself in a variety of ways in various people. After the data has been gathered and categorized, the subsequent step is to preprocess the text and tokenize it utilizing the tokenization method. In this process, the reader is breakdown into individual sub-words and assigned a unique ID to each subword. After preprocessing, the relevant features are extracted from the text and prepared as a feature vector for the model training. Machine learning or deep learning-based language models are used and trained in a supervised manner. Finally, the model's performance is evaluated using accuracy, recall, and F1 score measures. The detail of each step of the proposed technique is explained in the subsequent section, and an overview of the proposed pipeline is mentioned in the following figure.

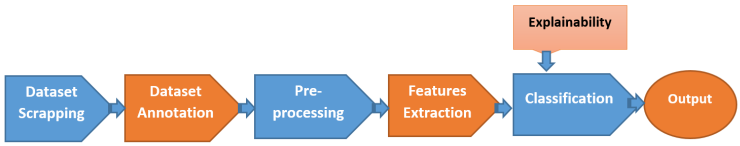

Fig. 1. Overview of proposed pipeline

3.1.1 Dataset Collection. We have used Apify, which is a scrapping platform. In our case, we had to scrap tweets, so we used the Twitter scrapper of Apify. We used the following lexicons and phrases/terms to search tweets for five types of depression (Bipolar depression, Atypical depression, Psychotic depression, Major depressive disorder, and Postpartum depression).

3.1.2 Lexicons Used for Scrapping Tweets. Major depressive disorder: "I have a major depressive disorder," "I am suffering from major depressive disorder," "I have major depression," "suffering from major depression." "Major depressive episode"

Bipolar disorder: "I have bipolar disorder," "suffering from bipolar disorder," "I have bipolar depression," "suffering from bipolar depression," "bipolar affective disorder," "bipolar mood disorder," "bipolar."

Atypical Depression: major depression with atypical features, atypical major depression, hypersomnia, feeling sad or hopeless, increased appetite" "weight gain, feeling worthless.

Psychotic depression: psychotic depression, delusional depression, psychotic depressive disorder, melancholic depression, "I have psychosis," "I have psychotic depression,"

Postpartum depression: "postbirth depression," "post-childbirth depression," "maternal depression," "I have postpartum depression."

Major Depressive Disorder: Major Depressive Disorder, is characterized by persistent sadness and hopelessness, as well as a lack of interest or pleasure in things. It is a significant condition that can disrupt an individual ability to carry out daily tasks and can affect both the emotional and physical well-being of the individuals

Bipolar Disorder: Another mental health disease known as bipolar disorder, which is characterized by spells of both depression and mania or hypomania. Bipolar disorder was once known as manic depression. It can impair a person's capacity to carry out daily activities and mood, energy, and activity levels.

Psychotic Depression: A person who suffers from psychotic depression, a subtype of major depressive disorder (MDD), has not only the symptoms associated with depression but also those related to psychosis.

Postpartum Depression: Postpartum depression (PPD), which is also called postnatal depression, is a depression type which can manifest itself in a new mother after she has given birth. It is a mental health disorder that affects a person's emotional and physical well-being and can interfere with their ability to care for themselves and their infant.

Atypical Depression: Atypical depression is categorized as a subtype of major depressive disorder (MDD) that characterizes the distinctive symptoms that separate it from other kinds of depression. Atypical depression is characterized by specific symptoms that distinguish it from other forms of depression. Because the symptoms of this type of depression are different from the typical symptoms of MDD, it is referred to as an "atypical" form of depression. After finalizing these search tags and keywords, I extracted the relevant tweets and saved them into Data Frame created with Pandas. The Data Frame was saved as a CSV file. This standard file format can be readily transferred to other programs and evaluated by other systems. I compiled this CSV file for the model training and tweets detection.

3.1.3 Dataset Annotation. Adding metadata or labels to a dataset to make it simpler to use and analyze within the context of machine learning and other data-driven applications is called dataset annotation. Adding information to data points inside a dataset, which may include descriptive labels, tags, or additional metadata, is called an annotation. Annotating a dataset is a crucial stage in preparing a dataset for use in machine learning applications. This step is critical because it offers the context and structure required to make the data useable for various tasks, including classification and natural language processing. Annotation is of utmost significance in supervised learning tasks in which a model is trained on labeled data.

This is because annotation supplies the model with the ground truth labels it utilizes to learn and make predictions. The type of data being annotated, as well as the tools and resources that are readily available, can determine whether the annotation process is carried out manually or automatically. Comparatively, automatic annotation uses algorithms and machine learning techniques to assign labels on established rules or patterns in the data. In contrast, manual annotation requires human annotators to evaluate the classify each data point in the dataset. We also followed the manual annotation of scraped tweets in this research study. The tweets are carefully labeled keeping in mind the context tweets which contain the word bipolar, atypical, or any other type of depression were not marked as depressed until and unless the context also matches the situation. Only those tweets where it is evident that a person is suffering from bipolar depression or any other kind of depression has been labelled as depressed. A sample image of Dataset Annotation with depression classes is mentioned in the figure 2. The remaining dataset annotation pic is attached in appendix.

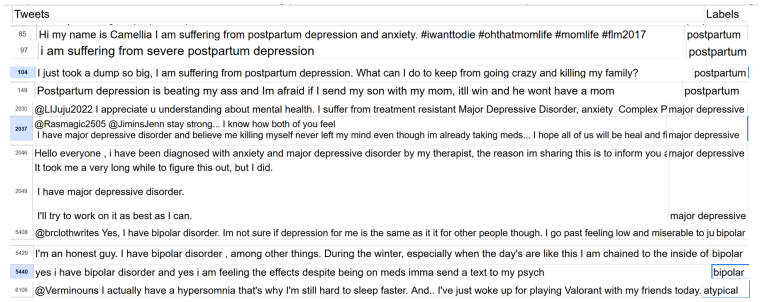

Fig. 2. Dataset

3.1.4 Data Preprocessing. The data preprocessing phase is an integral part of depression detection through tweets. The preprocessing step involves cleaning and preparing the Twitter data to be utilized for training a model. Removing noise and information useless from the tweets is also an essential part of data preprocessing. In this step, we deleted the URLs and hashtags and removed the tweets that contained profanity or inappropriate language. After this, we performed the Tokenization as an additional step for the data preparation. This process entails separating the text into its component subwords and providing a unique identifier for each subword. These were later split into training, validation, and testing for the deep learning model.

3.1.5 Features Extraction . The process of translating raw data into valuable features that may be used for tasks involving deep learning and natural language processing is called feature extraction. Put another way; it entails selecting and extracting the essential information from the input data and displaying it appropriately in a machine-learning model. Similarly, this process can entail several strategies, depending on the kind of data and the particular natural language tasks needed. When performing tasks involving natural language processing, feature extraction may require techniques such as bag-of-words or TF-IDF to express the text data in a format that is readable by computers, or any deep learning-based model can be accomplished for features extraction such as BERT. In this research, we used BERT for feature extraction. Natural language processing (NLP) endeavors can benefit significantly from the utilization of the sophisticated feature extractor that is BERT (Bidirectional Encoder Representations from Transformers), among other techniques such as Bag of Words (BoW), TF-IDF, and Glove. BERT is a pre-trained language model capable of learning rich and context-dependent text representations. Because of this capability, BERT is an excellent choice for feature extraction because it can be used to learn new words of text.

To use BERT as a feature extractor, we fed the preprocessed data as an input text through the pre-trained model and then acquired the contextualized embedding generated by the last layer of BERT. These embedding examples of a contextualized text representation are provided as input. Also, these embeddings that BERT caused can subsequently be used as features in downstream NLP tasks such as entity name recognition, question answering, and text classification; we used these embeddings for the type of depressions through tweets.

3.1.6 Depression Detection Using Machine Learning. A significant mental health condition that has a widespread global impact is depression. Effective therapy for depression depends on early detection and prompt action. Machine learning (ML) has shown promise in helping in the diagnosis of depression. Due to ML's capacity to discover patterns and associations in huge datasets, there has been an increase in interest in applying these techniques to diagnose depression

in recent years. In this research study, we explored many machine learning methods that have been applied to the identification of depression in this chapter, including Support Vector Machines, Random Forest, and Naive Bayes.

3.1.7 Depression Detection Using Deep Learning. The identification and diagnosis of depression in its earliest stages are of the utmost importance for its successful treatment. Deep learning (DL) strategies have demonstrated significant promise as a means of enhancing both the precision and effectiveness of depression diagnosis. The research study utilizes various deep learning methods such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Bidirectional Encoder Representations from Transformers (BERT). Out of these methods, the primary emphasis of this research study lies in assessing the performance of the BERT model for detecting depression and also investigated the concept of explainability with the BERT Prediction. The detail of the BERT model and Explainability is mentioned in the subsequent section

3.1.8 Tweets Classification Using BERT. The need for more available training data is a significant challenge in NLP. Most datasets which are specific to one task only comprise a few thousand to a few hundred thousand human training labels. This is attributed to the diverse nature of Natural Language Processing (NLP) as a field. In contrast, modern NLP models that rely on deep learning demonstrate advantages when exposed to extensive datasets, thriving particularly when trained on millions or even billions of annotated training examples. Researchers have put forth diverse approaches to develop general-purpose language representation models by leveraging vast quantities of unannotated text available on the internet. This process, known as pre-training, is intended to help close the current data gap. Compared to the training on smaller datasets from scratch, using a pre-trained model allows for significant improvements in accuracy when doing small data NLP tasks such as QA, classification and sentiment analysis. This can be fine-tuned using pre-trained data. BERT is being utilized mainly in the NLP domain for classification tasks.

BERT, "Bidirectional Encoder Representations from Transformers", is a pre-trained language model with state-of-the-art results in various natural language processing tasks, including text classification. BERT is credited with revolutionizing NLP. BERT a transformer-based model that can learn contextualized representations of words and sentences. It was trained on vast volumes of text data in an unsupervised manner while undergoing training, and it introduced the data in an unsupervised way. We can fine-tune the pre-trained model by applying it to a labeled dataset specific to the classification task we are attempting to solve with BERT before involving it in the problem we are trying to solve, as we used it for depression classification. When fine-tuning BERT, we need to add a classification layer on top of the model that has already been pre-trained and then train it using supervised learning on the dataset that has been labeled. The pre-trained BERT model and the classification layer are updated while the training is done to minimize the loss function.

When utilizing BERT for text classification, as in our case, the input text must first be broken down into sub-words, and then the input sequence must be supplemented by specialized tokens. we converted these tweets in preprocessing phase and then extracted the ticket through the tokenization process, which was used for the classification task. The tokens represent the entire input sequence, and the features extracted from its corresponding hidden representation can be incorporated into a classification model to make predictions regarding the final label. Figure 3 shows the architecture of Bert.

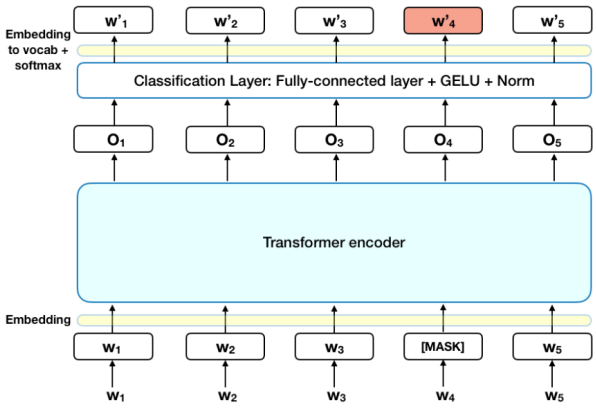

Fig. 3. Bert Language model architecture

To get complete insights into the task, we used the concept of explainability with BERT language modeling. The detail of the process is mentioned in a subsequent section.

3.1.9 Explainability. Explainable Artificial Intelligence (XAI), also known as XAI, is a methodology for developing deep learning models and systems open to human interpretation and can be explained to them. XAI's mission is to improve human understanding of how AI systems come to a decision or recommendations they do, to foster a greater level of trust and accountability. These goals are helpful to accomplish by providing insights into the decision-making process that AI systems use. The different artificial intelligence models, such as neural networks and BERT, can be explained using XAI approaches. Methods that are utilized frequently in XAI include the creation of feature importance scores, the visualization of decision boundaries, the generation of counterfactual explanations, and the provision of natural language explanations. XAI can be utilized with the BERT language model for the depression detection problem to promote transparency and accountability and provide insights into decision-making. Using XAI in conjunction with BERT can be accomplished in several ways. One of these ways is to construct attention maps, which call attention to the significant tokens in the input text responsible for the output classification. BERT generates attention maps for each ticket in the input sequence. These attention weights are based on the order in which the tickets appear in the series. These attention weights can be used to build attention maps emphasizing the significant words and phrases in the input text that played a role in classifying the data. The second approach, LIME Local Interpretable Model Agnostic Explanations), can also explain the classification decision. The complete diagram of the proposed pipeline is presented in figure 4.

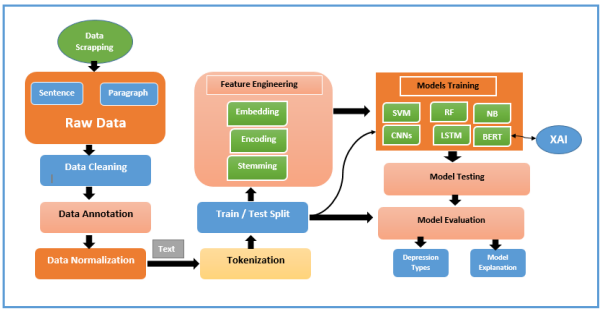

Fig. 4. An overview of proposed pipeline

3.2 Flowchart of the Proposed Solution

The flowchart of the proposed solution is given in figure 5.

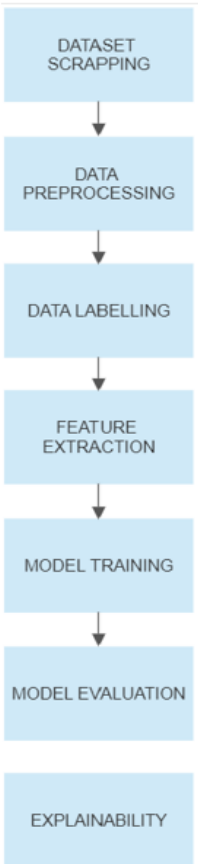

Fig. 5. Flowchart of proposed solution

3.2.1 Data Preprocessing. The following preprocessing techniques have been used in the dataset.

Text Normalization: We converted tweets to lowercase. Removed URLs or any web links which have no value in the text. Removed Twitter handles (e.g., @someone). Removed punctuation marks and numbers in the tweets.

Stopwords Removal: Removed words that do not have much importance, such as the, is, and etc.

3.2.2 Data Labelling. The tweets are labeled by reading the whole sentence context. If a tweet contains the word bipolar depression, I didn't label it to be bipolar because there is a high possibility that the tweet may be about a third person who is suffering from bipolar depression, so it was highly important to read the full sentence and then label the tweet. The tweets are only labeled as depressed if a first-person himself/herself is suffering from depression or has suffered from depression. All the tweets which did not fall under these conditions were labeled as not depressed. The dataset was verified by a domain expert i.e. psychiatrist.

3.2.3 Feature Extraction. We used BERT for feature extraction. Details of feature extraction were explained in Chapter 2.

3.2.4 Tokenization. Tokenization means to split or break the text into individual words. Tokenization is necessary because our machine learning model cannot understand long text, so we must convert text into some format that a model can process. We cannot do any NLP task without tokenization. In my case, I have used BERT as a tokenizer. The BERT tokenizer breaks a word into a subword, which helps to overcome the problem of out-of-vocabulary words. If a word is absent in the vocabulary, it is broken down into small chunks. In this way, the BERT model will now process this word as it will have the pre-trained embedding for that chunk portion of that word. In this way Out of Vocabulary problem (OOV) is resolved. After tokenization, padding is done to make all input text of equal length which is necessary. CLS token indicates the start of a sentence, and SEP indicates the end of a sentence. The input layer of the BERT model converts the token into a vector representation, also called embedding, which captures its meaning in the context of the sentence. We get a sequence of vectors in output where each vector corresponds to its input token text.

3.2.5 Attention Mechanism In BERT. Attention is used in BERT, which calculates the weight of each input text token. It helps as it helps to identify which token is more important in a sentence as compared to others. It assigns a weight or a score to each token, showing its importance.

3.2.6 Model Training. The approach involved utilizing a pre-trained BERT model, which was then fine-tuned using a dataset of tweets. BERT model was initially trained on a large corpus of text data. Below are some of the steps for model training using BERT.

The first step is tokenization. The input tweet text is broken down into words or subwords to overcome the issue of out-of-vocabulary words. After tokenization, we have to do formatting. BERT model takes input tokens of fixed length, so we have to ensure each sentence length is the same for this purpose, we add padding to ensure this step. We loaded a pre-trained model of BERT and fine-tune it to our cutom dataset. In other words, we can use a pre-trained model to fine tune to it our specific task in our case, task is sentiment analysis, so fine-tuning will require a less number of training data to give good results, which is a plus point as it will save a lot of our time and computational complexity.

3.2.7 Model Evaluation and Testing. After Fine-tuning the BERT model to our dataset, we have to evaluate the model on our validation dataset during training if the accuracy is not good on validation data we adjust the parameters also called hyperparameters until the loss is minimized and accuracy is increased. Finally, the fine-tuned BERT model is tested on our test dataset. We also used explainable AI to give the reasoning why the tweets were predicted as depressed.

3.3 Dataset and Implementation Details

There were 23,634 tweets and six classes in the dataset.

- Bipolar depression
- Psychotic depression
- Atypical depression
- Postpartum depression
- Major depressive disorder
- No Depression

Figure 6 illustrates the number of tweets in each class of the dataset.

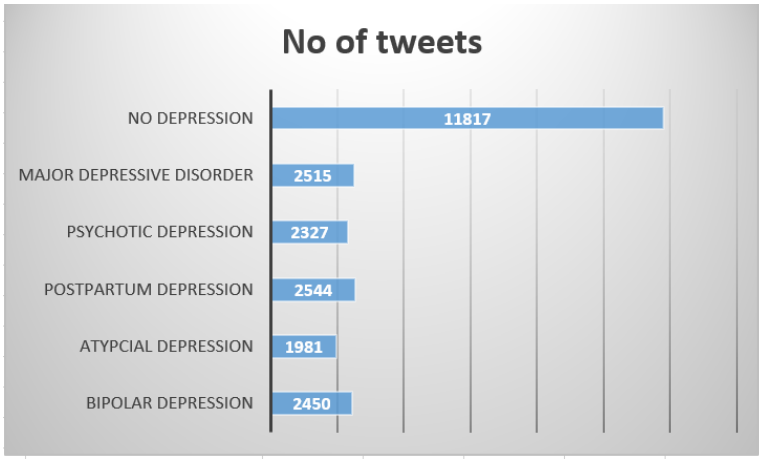

Fig. 6. Dataset Details

3.4 Libraries Used

The following libraries were used in the code.

- Pandas
- Regex
- NLTK
- Matplotlib
- Time
- Transformers
- PyTorch
- Ktrain
- Random
- Shap
- Numpy
- Tensorflow

We will briefly discuss why we used these libraries.

Pandas were used to read the information in the CSV file and load it into a Data Frame.Regex was used for preprocessing the dataset of tweets i.e. removing special characters, hashtags, mentions, stopwords, URLs, and emails that are noise in our dataset.NLTK was used for text processing i.e. tokenization. Matplotlib was used for data visualization i.e. bar charts, graphs, etc. Time library was used to analyze sentiments at multiple time intervals. Transformers was used because we have used BERT pre-trained model in my thesis and to use that we have to import the transformer's library. PyTorch was used to deal with sequential textual data of tweets. Ktrain was used to make the training process more robust. Random was used to shuffle tweets in the CSV file of the dataset. Shap was used for explainability to highlight which features in the tweet contributed in predicting the depression type. NumPy was used for data analysis. TensorFlow was used for training neural networks.

3.5 Programming Language

We used Python programming language. There are several reasons for it, which are given below.

- Open Source
- High Level Programming Language
- Executes code line by line, easy to resolve errors.
- No need to declare variables,
- Libraries are present which we can use and train our model instead of writing code from scratch.

4 EXPERIMENTAL EVALUATION

The purpose of conducting the research was to evaluate the capabilities of machine learning and deep learning models for the detection of depression using data collected from twitter. The results of our data analysis and model evaluation are presented in this chapter. This is followed by a discussion of the interpretation of the results, along with an examination of the practical implications, limitations, and potential avenues for future research. In this research, we used 23,634 tweets obtained from the Twitter platform of different individuals. These tweets were then subjected to preprocessing and feature engineering in order to extract information that was pertinent to depression identification. Using the collected features, We trained a number of machine learning models including SVM, Random Forest, and Naive Bayes. Additionally, deep learning models such as LSTM, CNN, and BERT were also used for depression detection. The performance analysis of each model was conducted using standard evaluation metrics such as precision, recall, accuracy, and F1-score. The research findings offered novel insights into the linguistic, semantic, and emotional aspects that contribute to the detection of depression. These findings also have a number of practical consequences for mental health practitioners, social media platforms, and individual users.

The detail of the experiments is explained in the below section with respect to the individual Machine Learning models and Deep learning models.

4.1 Depression Detection Using Machine Learning

Depression is an extremely significant mental disease that impacts the lives of millions of individuals all over the world. The effective treatment of a condition and the prevention of its long-term effects both begin with accurate diagnosis and prompt intervention. Platforms such as Twitter has access to a wealth of data that, when analyzed, can mental health and the emotions they are experiencing. In recent times, there has been a noticeable surge in the utilization of machine

learning algorithms to analyze data derived from social media platforms. These algorithms offer new prospects to detect depression at an early stage. Furthermore, we have conducted an assessment of the performance of widely recognized machine learning models, such as SVM, Random Forest, and Naïve Bayes, for the purpose of depression detection.

4.1.1 Depression Detection Using Support Vector Machine. Support Vector Machine (SVM) is a well-established supervised machine learning technique extensively applied in various classification problems. In the case of depression detection using data from Twitter, SVM is trained using a carefully chosen subset of the dataset, specifically curated for the training process. The algorithm is able to understand the patterns and links that exist between the attributes that are taken from the tweets and the labels that were associated with them. The SVM then makes use of these learned patterns in order to make predictions of unseen tweets (test data). According to the standard dataset division, the 70% data was used for training the model and the rest of the data was used for validation and test set. The different kernel and Gemma values are experimented with in order to achieve a better result. The performance evaluation of the model encompasses the assessment of accuracy, recall, precision, and F1-Score values. The corresponding precision, recall, F1-Score, and accuracy metrics for each type of depression are presented in the table.2. Similarly, overall precision recall F1 score and accuracy is given in table 3.

Table 2. Evaluating SVM Depression Classification Metrics

SVM	**Precision**	**Recall**	**F1 score**	**Support**
Atypical Depression	99	97	98	382
Bipolar Depression	90	93	92	484
Major Depression	83	85	84	496
Postpartum Depression	91	88	90	503
Psychotic Depression	87	85	86	494
No	99	99	99	2529

Table 3. Overall Accuracy Precision Recall and F1-score of SVM

SVM	**Precision**	**Recall**	**F1 score**	**Accuracy**
Atypical Depression	99	97	98	382

4.1.2 Depression Detection Using Random Forest. Random Forest is recognized as one of the most commonly employed machine learning algorithms, and it is well-known for its capacity to deal with high dimensional data as well as intricate correlations between features. RF is trained on a portion of the dataset that is divided according to the types of depression. The algorithm is able to understand the patterns and links that exist between the attributes that are taken from the tweets and the labels that are associated with them. The Random forest makes use of these learned patterns in order to speculate on the label of new tweets that have not been the part of training data. The Random Forest classifier is able to manage missing data and noisy features; it is considered to be an appropriate algorithm for depression detection. We also took advantage of this parameter and use Random forest for depression detection using the Twitter data set. The performance of the Random Forest (RF) model is assessed using metrics such as Precision,

Recall, F1-Score, and Support, which are presented in the table 4. Table 5 presents overall accuracy, precision, recall and F1 score values.

Table 4. Evaluating Random Forest Depression Classification Metrics

Random Forest	Precision	Recall	F1 score	Support
Atypical Depression	99	96	97	382
Bipolar Depression	95	92	94	484
Major Depression	89	81	85	496
Postpartum Depression	91	90	90	503
Psychotic Depression	85	89	87	494
No	98	100	99	2529

Table 5. Overall Accuracy Precision Recall and F1-score of Random Forest Classifier

Random Forest	Precision	Recall	F1 score	Accuracy
Overall Metrics Result	94.7	94.7	94.7	94.7

4.1.3 Depression Detection Using Naive Bayes. We also used naive Bayes algorithm to predict the type of depression in tweets. Naive Bayes first calculates the prior probability of each class in our case we have 6 classes so naive Bayes will calculate the prior probability of each class present in our dataset based on the frequency of labels. Once the prior probability is computed, the naive Bayes algorithm proceeds to calculate the probability of each feature across all five depression classes. The reason behind calculating probability of each feature in each class is that the algorithm understands the pattern that these words are coming in a particular class, so it helps when it will predict a test tweet in the dataset i.e. what type of depression is present in the test tweet of the dataset. It helps to calculate posterior probability in the dataset. Table 6 shows naive Bayes depression classification metrics and table 7 shows overall precision recall f1-score and accuracy using naive Bayes classification algorithm.

Table 6. Evaluating Naïve Bayes Depression Classification Metrics

Naive Bayes	Precision	Recall	F1 score	Support
Atypical Depression	98	91	94	382
Bipolar Depression	94	77	85	484
Major Depression	78	75	77	496
Postpartum Depression	86	89	87	503
Psychotic Depression	69	84	76	494
No	98	98	98	2529

Table 7. Overall Accuracy Precision Recall and F1-score of Naïve Bayes

Naive Bayes	Precision	Recall	F1 score	Accuracy
Overall Metrics Result	91.4	90.9	91.0	90.9

4.1.4 Machine Learning Classifiers Analysis. In conclusion, the decision on which classifier to use is depends on different factors. These factors include the characteristics of the data itself, the size of the data, and computational resources. These popular machine learning classifier SVM, RF, and NB each have pros and cons. As, SVM works well with high dimensional data that can be non-linearly separated, although it can be computationally expensive. Random Forest is resilient to noisy data and can account for missing values, but it has a risk of overfitting when it is applied to a highly correlated features set. Similarly, the naïve Bayes is both computationally efficient and able to deal with high dimensional data; nevertheless, it operated under the assumption that the characteristics are independent, which is not always the case. we considered all these classifiers and then performed a detailed analysis by individuals. The algorithms' performance was assessed by measuring accuracy, precision, recall, and F1-score metrics. For the depression detection analysis, Random Forest gave the best results. The detailed comparison of the classifier is mentioned in Table 8.

Table 8. Comparison of accuracy precision recall & F1 score of different machine learning classifiers

Machine Learning Classifiers	**Accuracy**	**Precision**	**Recall**	**F1 score**
SVM	94.4	94.5	94.4	94.4
Random Forest	94.7	94.7	94.7	94.7
Naive Bayes	90.9	91.4	90.9	91.0

4.2 Depression Detection Using Deep Learning

In recent times, there has been a significant surge in interest towards deep learning, primarily because of its ability to autonomously learn intricate features and patterns from vast amounts of data. One of the reasons for this is that it can detect depression. Deep learning methods, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and BERT have demonstrated promising results in the detection of depression across multiple modalities, such as speech, text, and images.

Deep learning possesses the capability to acquire representations of data at multiple levels of abstraction, enabling the model to capture intricate correlations between features. This is one of the advantages of utilizing deep learning to diagnose depression. Another benefit is the capability to manage data from a variety of modalities and integrate it all into a single model. However, in order to train, deep learning models may need a significant quantity of data as well as processing resources. These models may also be prone to overfitting if the dataset they are using is either too little or too unbalanced. Interpreting the outcomes of deep learning models can be challenging due to the fact that these models frequently incorporate a huge number of parameters and complex structures. Overall, deep learning shows promise for improving the accuracy and effectiveness of depression diagnosis; however, great consideration needs to be given to the construction of the model, the choice of data format, and the evaluation of the results. In this research, we considered CNNs, LSTM, and the BERT model of Deep learning to detect depression.

4.2.1 Depression Detection Using Convolutional Neural Networks (CNNs). CNN is used for the purposes of image identification, classification, and segmentation. CNN is an abbreviation for the phrase "convolutional neural network" CNNs are constructed by layering interconnected nodes that progressively learn to extract features from the input data. The structure and function of the visual cortex in the brain served as an inspiration for the development of CNNs. Convolutional layers, which carry out a sequence of mathematical operations to extract spatial characteristics from

the images that are fed into a CNN, are the fundamental elements that make up its architecture. In most cases, these layers are made up of a no of filters or kernels which are moved across the image while carrying out element-wise multiplication and addition in order to generate a feature map. Following this, the generated feature maps are sent through additional layers, such as pooling and activation layers, in order to further reduce the dimensionality of the model and boost its non-linearity. In this research, we considered the CNN where we used 4 Convolutional layers along batch normalization and pooling layers. To train the model the 10 epochs are used with 32 batch size and 10 epochs. The training process involved the utilization of cross-entropy loss and the Adam optimizer. The drop-out layer is used as the last layer with 6 as a parameter to identify the depression types. We trained the model with varying epochs and evaluated the model on the following metrics i.e. accuracy, recall, precision, and f1-score. The training loss, validation loss, training accuracy, and validation accuracy are reported in figure 7. Table 9 shows evaluation metric values for each type of depression, and table 10 shows overall evaluation metric values.

Table 9. Evaluating CNN's Depression Classification Metrics

CNN	Precision	Recall	F1 score	Support
Atypical Depression	94.5	94.4	94.4	382
Bipolar Depression	94.7	94.7	94.7	484
Major Depression	91.4	90.9	91	496
Postpartum Depression	77	85	81	503
Psychotic Depression	83	84	84	494
No	99	97	98	2529

Table 10. Average Evaluation Metrics for CNN Depression Classification

CNN	Precision	Recall	F1 score	Accuracy
Overall Metrics Result	92.9	92.6	92.7	92.6

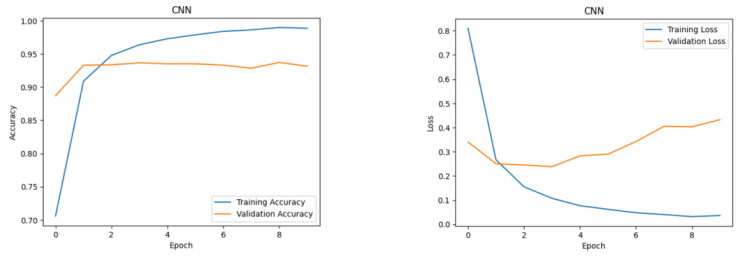

Fig. 7. CNN training validation accuracy and training validation loss

4.2.2 Depression Detection Using Convolutional Neural Networks (CNNs) with Glove. CNN combined with GloVe, which stands for "Global Vectors for Word Representation," is a method that is frequently used for text categorization and

sentiment analysis applications. GloVe is a well-known unsupervised approach that is used to generate word embedding. Word embedding is vector representations of words that exist in a space with a high dimension. These embeddings determine the semantic and syntactic linkages between words by analyzing their co-occurrence patterns in a vast body of text. These patterns are derived from the corpus of text. In the context of text classification, CNNs that use GloVe embedding are used to extract features from the text that is being fed into the network. These features are then fed into fully connected layers, where a decision is made on the classification. Typically, the GloVe embedding is pre-trained on huge corpora such as Wikipedia or Common Crawl, and then they are fine-tuned based on the particular task that is being performed. In this approach, we used Glove embedding as a concept of pertained model and embedded it with 3 layers' cnn followed by the max pooling layer. At layer 64 filters are used with 3*3 size and the ReLu activation function is used to make the model simpler. For depression classification, the dense layer is connected with the Softmax activation function. The model is trained with different hyperparameters such as "cross-entropy loss", Adam optimizer, batch size = 64, and 10 epochs. Accuracy(training and validation) and loss are illustrated in figure 8. Evaluation metrics for each type of depression and overall precision recall f1 score accuracy are given in table 11 and 12 respectively.

Figure 8 shows that as the epochs are increased, the training and validation loss decreases. On Contrary, the accuracy (training,validation) increases with the increase in number of epochs.

Table 11. Evaluating CNN with Glove Depression Classification Metrics

CNN with GloVe	Precision	Recall	F1 score	Support
Atypical Depression	86	90	88	696
Bipolar Depression	90	89	89	677
Major Depression	74	77	75	613
Postpartum Depression	86	90	88	743
Psychotic Depression	90	74	81	474
No	94	96	95	384

Table 12. Average Evaluation Metrics for CNN with GloVe Classification

CNN with GloVe	Precision	Recall	F1 score	Accuracy
Overall Metrics Result	87	86	86	86

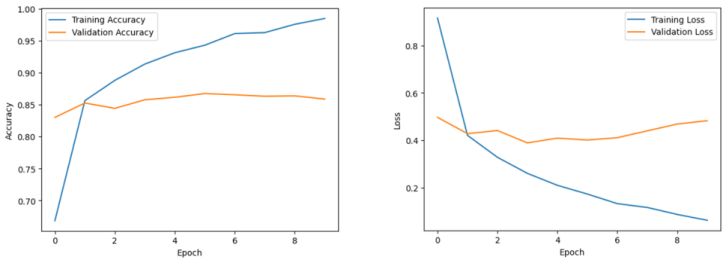

Fig. 8. CNN with GloVe training validation accuracy and training validation loss

4.2.3 Depression Detection Using LSTM. LSTM, an abbreviation for "Long Short-Term Memory," is a type of recurrent neural network (RNN) that is frequently utilized for tasks associated with natural language processing. These activities include language modeling, machine translation, and sentiment analysis. The problem of vanishing gradients in RNNs, which can cause the network to forget key information over lengthy sequences, is addressed by LSTMs, which were created specifically to manage this issue. In the context of depression detection, LSTMs may be used to model the temporal dependencies between the words in the input text and capture the emotional and cognitive patterns that are symptomatic of depression. This can be accomplished by modeling the temporal dependencies between the words in the input text. The input text is commonly represented as a string of word embedding, each of which is then fed into the LSTM layer one at a time. This process continues until the desired output is achieved. After that, the LSTM layer creates an output at each time step by updating its hidden state in accordance with the most recent input and the hidden state that came before it. In this analysis, I used 2 layers of the LSTM model with 64 LSTM units at each layer. The model is trained with different sets of hyperparameters such as drop out = 0.2, batch size = 32, loss- "cross-entropy", optimizer = "Adam", and epochs= 10. The evaluation metrics for each class in the dataset and overall evaluation metrics are given in table 13 and 14 respectively.

Table 13. Average Evaluation Metrics for LSTM Classification

LSTM	**Precision**	**Recall**	**F1 score**	**Support**
Atypical Depression	99	97	98	382
Bipolar Depression	93	90	92	484
Major Depression	85	85	85	496
Postpartum Depression	88	92	90	503
Psychotic Depression	86	85	85	494
No	99	99	99	2529

Table 14. Average Evaluation Metrics for LSTM Classification

LSTM	**Precision**	**Recall**	**F1 score**	**Accuracy**
Overall Metrics Result	94.6	94.5	94.5	94.5

Figure 9shows that as the epochs are increased, the training and validation loss decreases. On Contrary, the accuracy (training,validation) increases with the increase in number of epochs.

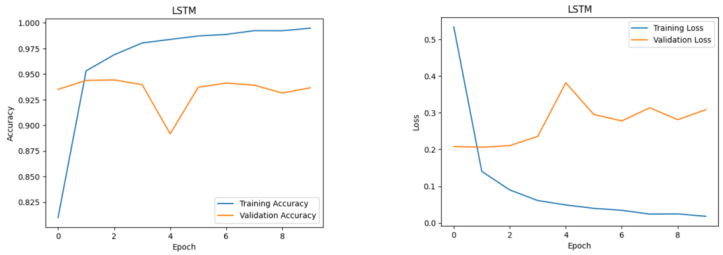

Fig. 9. LSTM training validation loss

4.2.4 Depression Detection Using LSTM with Glove. LSTM combined with GloVe, which stands for "Global Vectors for Word Representation," is a method that is frequently used for text categorization and sentiment analysis applications. GloVe is a well-known unsupervised approach that is used to generate word embedding. Word embedding is vector representations of words that exist in a space with a high dimension. These embeddings determine the semantic and syntactic linkages between words by analyzing their co-occurrence patterns in a vast body of text. These patterns are derived from the corpus of text. In the context of text classification, LSTMs that use GloVe embedding are used to extract features from the text that is being fed into the network. These features are then fed into fully connected layers, where a decision is made on the classification. Typically, the GloVe embedding is pre-trained on huge corpora such as Wikipedia or Common Crawl, and then they are fine-tuned based on the particular task that is being performed. In this approach, we used Glove embedding as a concept of pertained model.300 LSTM units were used with a dropout rate of 0.4. For depression classification, dense layer is connected with the Softmax activation function. The model is trained with different hyperparameters such as “cross-entropy loss”, Adamax optimizer, batch size = 32, and 10 epochs. The evaluation metrics for each class are shown in table 15 and for all classes evaluation metrics are shown in table 16.

Table 15. Evaluating LSTM with GloVe Depression Classification Metrics

LSTM with GloVe	**Precision**	**Recall**	**F1 score**	**Support**
Atypical Depression	94	96	95	382
Bipolar Depression	0	0	0	484
Major Depression	27	91	41	496
Postpartum Depression	46	23	31	503
Psychotic Depression	0	0	0	494
No	98	99	99	2529

Table 16. Average Evaluation Metrics for LSTM with GloVe Classification

LSTM with GloVe	**Precision**	**Recall**	**F1 score**	**Accuracy**
Overall Metrics Result	65.5	70.4	65.8	70.4

The training validation accuracy and training and validation loss is shown in figure 10 shows that as the epochs are increased, the training and validation loss decreases. On Contrary, the accuracy (training, validation) increases with the increase in number of epochs.

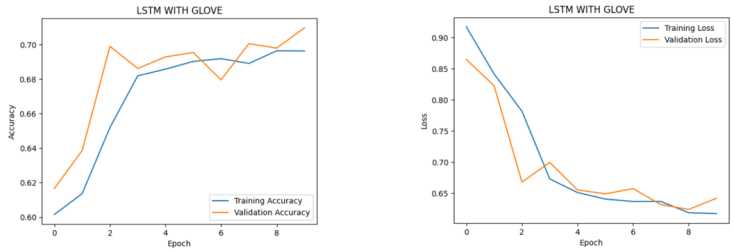

Fig. 10. LSTM with GloVe training validation accuracy and training validation loss

4.2.5 Depression Detection Using BERT. BERT, a pre-trained language model, has demonstrated state-of-the-art performance across various natural language processing tasks, including sentiment analysis, text classification, and question answering. The BERT is built on transformer architecture, and it is able to recognize the context in which words are employed and the intricate relationships that exist between the words in the phrase. For depression detection, BERT is used to encode the data into embedding's. BERT can automatically learn the encoded meaning and context of the input data, and it does not require considerable amount of feature engineering or domain specific knowledge. We fine-tuned BERT on our dataset. Bert gave the best results as the accuracy score came 0.96. It is because BERT is trained on huge corpus of text data. Similarly, BERT also handles the problem of out of vocabulary words.

Table 17 presents the precision, recall, F1-score, and support metrics for each type of depression. In table 18, the combined results are displayed. BERT achieves the highest precision, recall, F1-score, and accuracy values compared to the other models utilized in the study.

Table 17. Evaluating Bert Depression Classification Metrics

Bert	**Precision**	**Recall**	**F1 score**	**Support**
Atypical Depression	100	99	100	199
Bipolar Depression	96	95	96	238
Major Depression	87	88	88	242
Postpartum Depression	95	93	94	261
Psychotic Depression	89	90	90	233
No	99	99	99	1271

Table 18. Average Evaluation Metrics for LSTM with GloVe Classification

Bert	**Precision**	**Recall**	**F1 score**	**Accuracy**
Overall Metrics Result	96	96	96	96

4.2.6 Deep Learning Classifiers Analysis. In conclusion, the decision on which classifier to use is determined by several different aspects. These factors include the characteristics of the data itself, the size of the data, and computational resources. These popular deep learning classifiers CNN, CNN with Glove, LSTM, LSTM with Glove, and Bert all have their pros and cons. we considered all these classifiers and then performed a detailed analysis by individuals. The evaluation metrics were accuracy, precision, recall, and f1-score. For the depression detection analysis, we found BERT as the best method. The detailed comparison of the classifier is mentioned in table 19.

Table 19. Comparison of accuracy precision recall & F1 score of different Deep Learning Models

Deep Learning Models	Accuracy	Precision	Recall	F1 score
CNN	93	93	93	93
CNN with GloVe	86	87	86	86
LSTM	94	95	94	94
LSTM with GloVe	70	66	70	66
Bert	96	96	96	96

4.3 Model Explainability

Explainable Artificial Intelligence is a burgeoning subfield in AI that strives to develop models that can explain their decision-making process and provide the model's detailed insights. The mission of XAI is to find a way to bridge a gap between the AI models and requirements for human interpretability, accountability, and openness. The need for XAI originates from the fact that many complicated models can be difficult to read and explain even for professionals that specialize in the relevant discipline. This can result in a lack of trust in AI systems can limit their use in different application areas such as the healthcare industry, the financial sector and the legal system so on. In this research, we also employed the concept of XAI's to find the insights of the BERT model, and how the depression class is predicted, which are the main keywords that the model is being used to assign a specific type of depression on the basis of the Twitter data. Though there are different techniques to implement the XAI's in the model such as pre-model explainability, model explainability, and post-model explainability. Here we used the post-model-explain ability with the help of shapely, a well-known Python library for explainability. We have included a few examples of each type of depression with model highlighting the words in the sentence which indicate that particular type of depression.

In figure 11 as shown below the model has correctly predicted the tweet is referring to bipolar depression. The model has highlighted the words bipolar and disorder in green colour which indicates that these two words are emphasizing bipolar depression in the tweet.

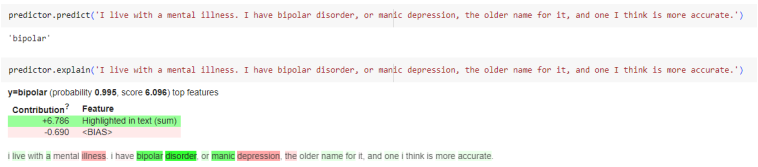

Fig. 11. explainability for bipolar depression

The results of explainability of other types of depression classes are included in the appendix.

5 LIMITATIONS AND FUTURE WORK

Depression detection is such a complicated and multi-faceted disorder; it can be challenging to accurately capture all of its facets in a single dataset. The data that is utilized to train depression detection models must be high quality and sufficiently representative for those models to produce accurate results. In addition, there is a lack of limited availability of label data, which also tried to overcome in this research study. As we scrapped the depression database of Twitter and made it public for the research community. Still, various types of other depression are not addressed in this corpus. Other, concerns of a moral and legal nature are raised in connection with the utilization of depression detection models. These models carry with them the potential to be utilized in a manner that stigmatizes and discriminates against those who suffer from depression. There are also concerns regarding the privacy and security of sensitive health data, particularly of the models utilized on social media or other public datasets.

In a further study, the primary focus may shift to the development of datasets that are both more representative and diverse, as well as the development of techniques that can identify and address the biases in the model. Depression often occurs in conjunction with other conditions such as anxiety. The deep learning techniques for the detection and diagnosis of these disorders in conjunction with depression could be investigated in future research. Since depression can present itself in a variety of ways, it may be beneficial to incorporate several modalities of data in addition to textual information. Some examples of these modalities are audio, video, and physiological signals. In subsequent research, the detection of depression using multimodal data could be investigated.

In conclusion, the application of machine learning and deep learning models for the identification of depression demonstrated encouraging results. In this research, we considered SVM, Random Forest, naïve Bayes, CNN, CNN with GloVe embedding, LSTM, LSTM with GloVe embedding, and BERT model. According to the research findings that I obtained, deep learning models LSTM and BERT outperformed. Among LSTM and BERT, the BERT performed best. Overall, the findings of the research work demonstrated the promise of machine and deep learning approaches for the detection of depression and underline the need for future research to develop more accurate, and efficient models.

6 ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my research advisor, Dr Waseem Shahzad whose extensive knowledge and constant guidance made it possible for me to conduct this study and present my findings. His advice and constructive feedback has been instrumental in shaping this work. I am also extremely grateful to Umair Arshad at National University of Computer and Emerging Sciences, he provided me with invaluable insights and suggestions throughout this project. His constructive criticism and constant encouragement served as motivation to improve my work and make it what it is today. I would like to thank Muhammad Farrukh Bashir at Riphah International University, who helped me in finding students of psychology for the labelling of the dataset. I would also like to thank Dr Ibad-ul-haq (Consultant Psychiatrist) who verified labelling of the dataset voluntarily. Lastly, I would thank my family for their constant support and understanding throughout this challenging project. They have been our source of inspiration and motivation, providing the strength I needed to keep pushing forward. This work is a culmination of many people's efforts, and for this, I am forever grateful. To everyone who contributed in one way or another, I sincerely say, Thank You.

7 FUNDING

There were no external funding used for the research.

8 CODE AVAILABILITY

The developed code will be made available through github repository.

REFERENCES

[1] Priyanka Arora and Parul Arora. 2019. Mining twitter data for depression detection. In *2019 international conference on signal processing and communication (ICSC)*. IEEE, 186–189.

[2] KS Aswathy, PC Rafeeque, and Reena Murali. 2019. Deep learning approach for the detection of depression in twitter. In *Proceedings of the International Conference on Systems, Energy Environment (ICSEE)*.

[3] Luca Bacco, Andrea Cimino, Felice Dell'Orletta, and Mario Merone. 2021. Explainable sentiment analysis: a hierarchical transformer-based extractive summarization approach. *Electronics* 10, 18 (2021), 2195.

[4] Bara Bataineh, Rehab Duwairi, and Malak Abdullah. 2019. ArDep: an Arabic lexicon for detecting depression. In *Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*. 146–151.

[5] Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media*. 75–84.

[6] Junyeop Cha, Seoyun Kim, and Eunil Park. 2022. A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community. *Humanities and Social Sciences Communications* 9, 1 (2022), 1–10.

[7] Xuetong Chen, Martin Sykora, Thomas Jackson, Suzanne Elayan, and Fehmidah Munir. 2018. Tweeting your mental health: An exploration of different classifiers and features with emotional signals in identifying mental health conditions. (2018).

[8] Paula Glenda Ferrer Cheng, Roann Munoz Ramos, Jó Ágila Bitsch, Stephan Michael Jonas, Tim Ix, Portia Lynn Quetulio See, and Klaus Wehrle. 2016. Psychologist in a pocket: lexicon development and content validation of a mobile-based app for depression screening. *JMIR mHealth and uHealth* 4, 3 (2016), e5284.

[9] Ginetta Collo and Emilio Merlo Pich. 2018. Ketamine enhances structural plasticity in human dopaminergic neurons: possible relevance for treatment-resistant depression. *Neural regeneration research* 13, 4 (2018), 645.

[10] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*. 51–60.

[11] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*. 31–39.

[12] Sally C. Curtin. 2020. State suicide rates among adolescents and young adults aged 10–24: United States, 2000–2018. https://stacks.cdc.gov/view/cdc/93667 Accessed: 2023-06-03.

[13] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, Vol. 7. 128–137.

[14] Bruce Ferwerda and Marko Tkalcic. 2018. You are what you post: What the content of Instagram pictures tells about users' personality. In *The 23rd International on Intelligent User Interfaces, March 7-11, Tokyo, Japan*. CEUR-WS.

[15] Tapotosh Ghosh, Md Hasan Al Banna, Md Jaber Al Nahian, Mohammed Nasir Uddin, M Shamim Kaiser, and Mufti Mahmud. 2023. An attention-based hybrid architecture with explainability for depressive social media text detection in Bangla. *Expert Systems with Applications* 213 (2023), 119007.

[16] Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C Eichstaedt, and Lyle H Ungar. 2019. What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the international AAAI conference on web and social media*, Vol. 13. 236–246.

[17] Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. *arXiv preprint arXiv:2209.07494* (2022).

[18] MA Islam, SD Barna, H Raihan, and M Khan. 2020. NA, and Hossain MT (2020) Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: A web-based crosssectional survey. *PLoS One* 15, 8 (2020), e0238162.

[19] Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. 2018. Depression detection from social network data using machine learning techniques. *Health information science and systems* 6 (2018), 1–12.

[20] Fazida Karim, Azeezat A Oyewande, Lamis F Abdalla, Reem Chaudhry Ehsanullah, and Safeera Khan. 2020. Social media use and its connection to mental health: a systematic review. *Cureus* 12, 6 (2020).

[21] Aleksandar Kecojevic, Corey H Basch, Marianne Sullivan, and Nicole K Davi. 2020. The impact of the COVID-19 epidemic on mental health of undergraduate students in New Jersey, cross-sectional study. *PloS one* 15, 9 (2020), e0239696.

[22] Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific reports* 10, 1 (2020), 1–6.

[23] Harnain Kour and Manoj Kumar Gupta. 2022. Depression and Suicide Prediction Using Natural Language Processing and Machine Learning. In *Cyber Security, Privacy and Networking: Proceedings of ICSPN 2021*. Springer, 117–128.

[24] Abhay Kumar, Vaibhav Pratihar, Sheshank Kumar, and Kumar Abhishek. 2021. Predicting Depression by Analysing User Tweets. In *Machine Vision and Augmented Intelligence—Theory and Applications: Select Proceedings of MAI 2021*. Springer, 633–644.

[25] Victor Leiva and Ana Freire. 2017. Towards suicide prevention: early detection of depression on social media. In *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4*. Springer, 428–436.

[26] Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. Sensemood: depression detection on social media. In *Proceedings of the 2020 international conference on multimedia retrieval*. 407–411.

[27] Ammara Nusrat, Hamza Farooq Gabriel, Sajjad Haider, Shakil Ahmad, Muhammad Shahid, and Saad Ahmed Jamal. 2020. Application of Machine Learning Techniques to Delineate Homogeneous Climate Zones in River Basins of Pakistan for Hydro-Climatic Change Impact Studies. *Applied Sciences* 10, 19 (2020). https://doi.org/10.3390/app10196878

[28] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*. 88–97.

[29] Rafał Poświata and Michał Perełkiewicz. 2022. OPI@ LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text using RoBERTa Pre-trained Language Models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 276–282.

[30] Esteban A Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. 2020. A dataset for research on depression in social media. In *Proceedings of the 28th ACM conference on user modeling, adaptation and personalization*. 338–342.

[31] Lawrence Robinson and Melinda Smith. 2023. The role social media plays in mental health. https://www.helpguide.org/articles/mental-health/social-media-and-mental-health.htm Accessed: 2023-06-03.

[32] Renata L Rosa, Demóstenes Z Rodríguez, Gisele M Schwartz, Ivana de Campos Ribeiro, and Graça Bressan. 2016. Monitoring system for potential users with depression using sentiment analysis. In *2016 IEEE International Conference on Consumer Electronics (ICCE)*. Ieee, 381–382.

[33] Ramin Safa, Peyman Bayat, and Leila Moghtader. 2022. Automatic detection of depression symptoms in twitter using multimodal analysis. *The Journal of Supercomputing* 78, 4 (2022), 4709–4744.

[34] Muskaan Singh and Petr Motlicek. 2022. IDIAP Submission@ LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 362–368.

[35] Sprout Social. 2023. 50+ of the most important social media marketing statistics for 2023. https://sproutsocial.com/insights/social-media-statistics/ Accessed: 2023-06-03.

[36] Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access* 7 (2019), 44883–44893.

[37] Dylan Walsh. 2022. Social media use linked to decline in mental health. https://mitsloan.mit.edu/ideas-made-to-matter/study-social-media-use-linked-to-decline-mental-health Accessed: 2023-06-03.

[38] Jianlong Zhou, Hamad Zogan, Shuiqiao Yang, Shoaib Jameel, Guandong Xu, and Fang Chen. 2021. Detecting community depression dynamics due to covid-19 pandemic in australia. *IEEE Transactions on Computational Social Systems* 8, 4 (2021), 982–991.

[39] Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web* 25, 1 (2022), 281–304.

[40] Chiara Zucco, Barbara Calabrese, and Mario Cannataro. 2017. Sentiment analysis and affective computing for depression monitoring. In *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 1988–1995.

[41] Chiara Zucco, Huizhi Liang, Giuseppe Di Fatta, and Mario Cannataro. 2018. Explainable sentiment analysis with applications in medicine. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1740–1747.

9 APPENDICES

9.1 Model Explainability For Remaining Depression Classes

In figure 12 as shown below the model has correctly predicted the tweet indicates atypical depression. The model has highlighted the words hypersomnia and if you read the whole sentence we can see that the person doing the tweet is suffering from hypersomnia. Hypersomnia is a symptom of atypical depression so this person doing the tweet is suffering from atypical depression.

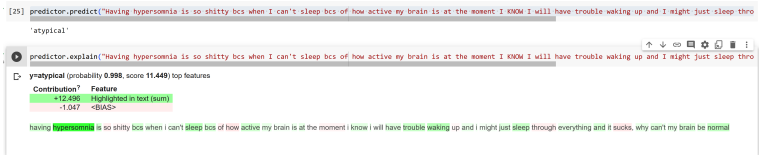

Fig. 12. explainability for atypical depression

As shown in figure 13, the model has correctly predicted the tweet indicates psychotic depression. The model has highlighted the words psychotic and hopeless. If a person is feeling worthless and hopeless then there is a high chance that person is suffering from psychotic depression.

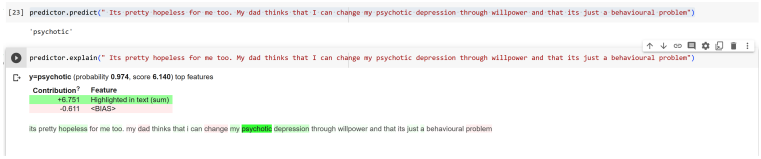

Fig. 13. explainability for psychotic depression

In figure 14, the model has correctly predicted that in the tweet mother is suffering from postpartum depression. Postpartum depression is mainly in mothers who are pregnant or have just given birth to their new child. The model has highlighted the words postpartum depression and mothers by looking at these words

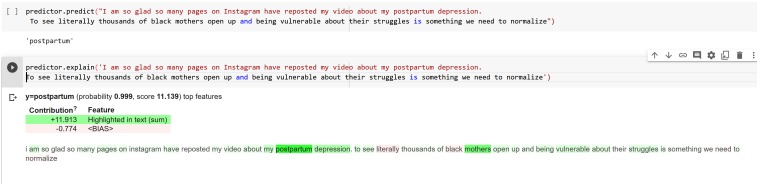

Fig. 14. explainability for postpartum depression

In the following figure, the person doing the tweet was suffering from major depression the model has correctly predicted it and by using explainable AI it has highlighted the words major depressive disorder. This means that the model made prediction by looking at these words.

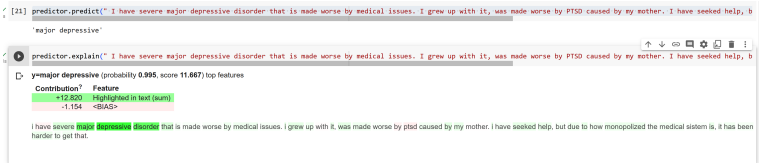

Fig. 15. explainability for major depression

In the following figure, the model predicted the tweet has no depression the person doing the tweet is not suffering from depression. The model made the decision by highlighting the words in the tweet which are in green colour.

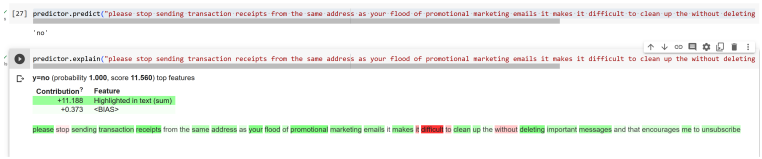

Fig. 16. explainability for no depression

9.2 Dataset Annotation

In the below figure, we can see few examples of the dataset annotation for atypical depression, psychotic depression and no depression class.

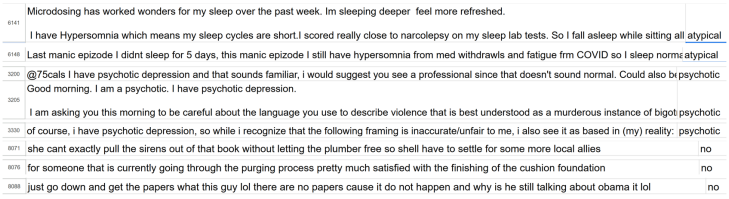

Fig. 17. Dataset Part b

Received 19 June 2023; revised ...; accepted ...